Appeared in:
Lynette Hirschmann, Martin Krallinger, and Alfonso Valencia, editors.
Proceedings of the Second BioCreative Challenge Evaluation Workshop.
Centro Nacional de Investigaciones Oncologicas, CNIO, Madrid, Spain, 2007.
pp. 89–92

Named Entity Recognition with Combinations of Conditional Random Fields

Roman Klinger
roman.klinger@scai.fhg.de

Christoph M. Friedrich
christoph.friedrich@scai.fhg.de

Juliane Fluck
juliane.fluck@scai.fhg.de

Martin Hofmann-Apitius
martin.hofmann-apitius@scai.fhg.de

Fraunhofer Institute for Algorithms and Scientific Computing (SCAI)
Department of Bioinformatics
Schloss Birlinghoven, 53754 Sankt Augustin, Germany

Abstract

The *Gene Mention* task is a *Named Entity Recognition* (NER) task for labeling gene and gene product names in biomedical text. To deal with acceptable alternatives additionally to the gold standard, we use combinations of *Conditional Random Fields* (CRF) together with a normalizing tagger. This process is followed by a postprocessing step including an acronym disambiguation based on Latent Semantic Analysis (LSA). For robust model selection we apply 50-fold *Bootstrapping* to obtain an average F-Score of 84.58 % on the trainingset and 86.33 % on the test set.

Keywords: named entity recognition, text mining, data mining, conditional random fields, multi model approach

1 Introduction

In general, machine-learning solutions deal with a single truth. One characteristic in BioCreative 2006 compared to common NER tasks is that the training data contains acceptable alternatives for gene and protein names next to the gold standard. One problem using only the gold standard is that this information is possibly more ambiguous than necessary. For example in the sentence "*On the other hand factor IX activity is decreased in coumarin treatment with factor IX antigen remaining normal.*" the gold standard is the twice annotation of *factor IX*. The alternative annotation gives the information that finding *factor IX antigen* is just as well. But in "*The arginyl peptide bonds that are cleaved in the conversion of human factor IX to factor IXa by factor XIa were identified as Arg145-Ala146 and Arg180-Val181.*" the gold standard is finding *human factor IX* and *factor IXa* and *factor XIa* but the alternative gives us the possibility of *factor IX* instead of *human factor IX*.

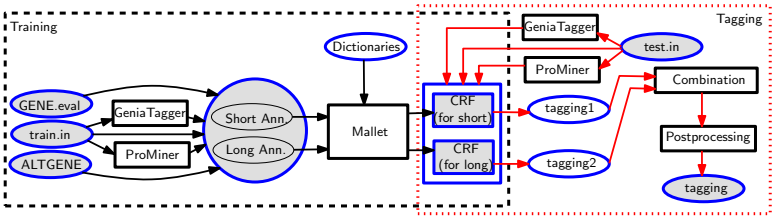

Figure 1: Workflow of our system.

We address that problem with a multi model approach using the Conditional Random Fields [5] implementation *Mallet* [7] which showed superior results in BioCreative 2004 [4] and our previous works [2].

2 System Description

The developed system is inspired by [8, 9]. A sketch of the workflow can be found in figure 1. At first the external tools *GeniaTagger* [11] and *ProMiner* [3] are called. Their results are used as IOB-features, which form the input for *Mallet* to build multiple Conditional Random Fields together with the sentences (in the file train.in) and the annotation information (in files GENE.eval and ALTGENE.eval).

Table 1: Strategies to combine different annotations. For the examples let us assume to have *fibrinogen degradation products* as annotation from the model trained on long annotations and *fibrinogen* and *FDP* as annotations from the model trained on short annotations on the text part *fibrinogen degredation products (FDP)*.

No.	Strategy	Example
1	Use long annotation first, then add short annotation (without overlaps)	*fibrinogen degradation products* ; *FDP*
2	Use short annotation first, then add long annotation (without overlaps)	*fibrinogen* ; *FDP*
3	Greedy: Combine both (with overlaps)	*fibrinogen* ; *FDP* ; *fibrinogen degradation products*

These multiple models deal with mentioned ambiguities by building one annotation out of the shortest possibilities and one out of the longest ones, each without overlaps. In the first example sentence mentioned in the introduction the short annotations are the ones from the gold standard, but in the second sentence we would use *factor IX* instead of *human factor IX*.

The generated models can then be used for tagging the new sentences (here we assume them to be in the file test.in) followed by the combination of these different outputs. We tried three different methods displayed in table 1. In the example in the second method nothing is added because the long annotation overlaps the short annotation.

The last step is postprocessing. Unequal numbers of closing or opening brackets are corrected and an acronym disambiguation using latent semantic analysis is conducted. It concerns the high frequent ambiguous acronyms *CAD*, *CSF*, *REM* or *CAP*. This concept study works here only at the sentence level but can be shown to be more powerful, if the full sentence context will be available.

3 Analysis and Results

For selecting the training parameters of the conditional random fields we use bootstrapping [1] with 50 replicates having approximatly 9480 training and 5520 validation examples in every replicate.

In our rich feature set we have different types of features like morphological [8] (some automatically generated [10]), dictionaries ([6] and self-made), offset conjunction and part-of-speech/shallow parsing information from the *GeniaTagger* [11]. Additionally we use the tagging information of the *ProMiner* [3], which achieves a precision of 0.88 on the training and 0.87 on the test set but a lower recall.

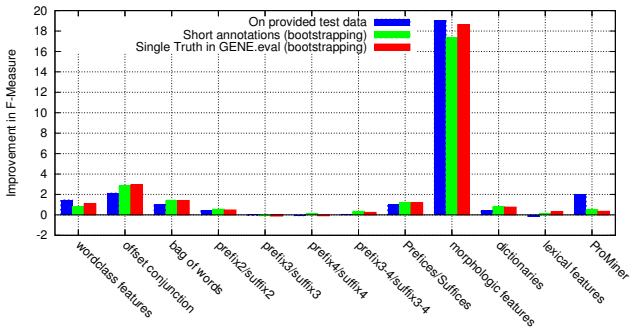

Figure 2: Influence of features estimated by omitting them.

Two classes of parameters are most important: The combination and selection of features and the tokenisation of the text. The impact of each feature or group of features was computed by building a conditional random field without them. It is displayed in figure 2. Morphological features are of overwhelming importance followed by offset conjunction. Interestingly the ProMiner has only a minor impact on the training set but improves results on the test set (with 2%). So it can be concluded that the performed approximative search has a higher impact than the simple dictionary matching.

The improvement of combinations of features is complex as can be seen in the case of prefixes. We detect an higher importance of using prefixes and suffixes of all lengthes (2–4) in comparison to only using these with length 2. This is not expectable because prefixes and suffixes of length 3 and 4 have no impact. This is an example for features not being independent as can also be seen in figure 2. It is not possible to have a greedy analysis of all combinations of the features because of prohibitive training times (about 1–2 hours, depending on the size of the feature set). Instead we conducted a systematic feature analysis based on attribute groups.

Analysing the tokenisation, we started with a complex tokenisation (inspired by [9]) reaching a mean F-score of 0.821 (using bootstrapping) on the system trained on the gold standard information in GENE.eval.

Table 2: Results on the trainingset (averaged over 50 bootstrap replicates) and on the test set after postprocessing and disambiguation (Standard deviation is given in brackets, submitted runs are marked with a *).

	Bootstrapping on Trainingset						On Testset		
Model	Precision		Recall		F-Score		Precision	Recall	F-Score
Long	86.30	(0.0065)	79.53	(0.0094)	82.78	(0.0064)	87.41	80.29	83.70
Short*	**86.87**	(0.0054)	81.94	(0.0106)	84.33	(0.0069)	**88.57**	83.83	86.13
Greedy*	80.21	(0.0069)	**89.47**	(0.0057)	**84.58**	(0.0047)	82.02	**90.63**	86.11
Long first*	85.38	(0.0060)	83.63	(0.0079)	84.50	(0.0055)	87.27	85.41	**86.33**
Short first	83.83	(0.0063)	84.81	(0.0065)	84.32	(0.0048)	85.50	85.61	85.56
GENE.eval	86.61	(0.0071)	81.76	(0.0123)	84.11	(0.0076)	87.86	83.53	85.64

Splitting always on dashes improves the results to 0.835.

The results of our systems on the training set and on the test set are displayed in table 2. The ratios between the experiments on the test data and on the training data with different models are similar, so we can assume bootstrapping as an appropriate choice for model selection. We see that it is already useful only to select a special subset of the alternatives for training: The annotation made by the short expert yields in better results than the one in GENE.eval. The combination of the short and long ones has further impact dependent on the strategie: Adding the short and long annotation the greedy way yields in a very high recall but a lower precision than the other methods. Using the long annotation and adding the short one gives us an higher precision and the highest F-Score of the different strategies.

References

[1] Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Chapman & Hall/CRC, 1993.

[2] C. M. Friedrich, T. Revillion, M. Hofmann, and J. Fluck. Biomedical and chemical named entity recognition with conditional random fields: The advantage of dictionary features. In S. Ananiadou and J. Fluck, editors, *Proceedings of the Second International Symposium on Semantic Mining in Biomedicine (SMBM 2006)*, pages 85–89, 2006.

[3] D. Hanisch, K. Fundel, H. T Mevissen., R. Zimmer, and J. Fluck. ProMiner: Organism-specific protein name detection using approximate string matching. In *Proceedings of the BioCreative Challenge Evaluation Workshop 2004*, 2004.

[4] Lynette Hirschman, Alexander Yeh, Christian Blaschke, and Alfonso Valencia. Overview of biocreative: critical assessment of information extraction for biology. *BMC Bioinformatics*, 6 Suppl 1:S1, 2005.

[5] J.D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001)*, pages 282–289. Morgan Kaufmann Publishers, 2001.

[6] Kevin Lerman, Yang Jin, Eric Pancoast, and Ryan McDonald. Biotagger. Software.

[7] A. K. McCallum. MALLET: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu, 2002.

[8] R. McDonald and F. Pereira. Identifying gene and protein mentions in text using conditional random fields. *BMC Bioinformatics*, 6 (Suppl 1)(S6), 2005.

[9] B. Settles. ABNER: an open source tool for automatically tagging genes, proteins and other entity names in text. *Bioinformatics*, 21(14):3191–3192, 2005.

[10] Burr Settles. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the COLING 2004 International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPBA)*, Geneva, Switzerland, 2004.

[11] Yoshimasa Tsuruoka, Yuka Tateishi, Jin-Dong Kim, Tomoko Ohta, John McNaught, Sophia Ananiadou, and Jun'ichi Tsujii. Developing a robust part-of-speech tagger for biomedical text. In *Advances in Informatics - 10th Panhellenic Conference on Informatics, LNCS 3746*, pages 382–392, 2005.

Acknowledgements

This work has been partially funded by the MPG-FhG Machine Learning Collaboration
`http://lip.fml.tuebingen.mpg.de/`